ILLUSTRATIVE EXAMPLES OF PRINCIPAL COMPONENT ANALYSIS USING SPSS-X/FACTOR*

W. T. Federer, C. E. McCulloch and N. J. Miles-McDermott

BU-928-M November 1986

ABSTRACT

In order to provide a deeper understanding of the workings of principal components, four data sets were constructed by taking linear combinations of values of two uncorrelated variables to form the X-variates for the principal component analysis. The examples highlight some of the properties and limitations of principal component analysis.

This is part of a continuing project that produces annotated computer output for principal component analysis. The complete project will involve processing four examples on SAS/PRINCOMP, BMDP/4M, SPSS-X/FACTOR, GENSTAT / PCP, and SYSTAT / FACTOR. We show here the results from SPSS-X/FACTOR, Release 2.2.

---

* Supported by the U.S. Army Research Office through the Mathematical Sciences Institute of Cornell University.

1. INTRODUCTION

Principal components is a form of multivariate statistical analysis and is one method of studying the correlation or covariance structure in a set of measurements on m variables for n observations. For example, a data set may consist of n = 260 samples and m = 15 different fatty acid variables. It may be advantageous to study the structure of the 15 fatty acid variables since some or all of the variables may be measuring the same response. One simple method of studying the correlation structure is to compute the m(m-1)/2 pairwise correlations and note which correlations are close to unity. When a group of variables are all highly inter-correlated, one may be selected for use and the others discarded or the sum of all the variables may be used. When the structure is more complex, the method of principal components analysis (PCA) becomes useful.

In order to use and interpret a principal component analysis, there needs to be some practical meaning associated with the various principal components. In Section 2 we describe the basic features of principal components and in Section 3 we examine some constructed examples using SPSS-X/FACTOR to illustrate the interpretations that are possible. In Section 4 we summarize our results.

2. BASIC FEATURES OF PRINCIPAL COMPONENT ANALYSIS

PCA can be performed on either the variances and covariances among the m variables or their correlations. One should always

check which is being used in a particular computer package program. (SPSS-X can only carry out a PCA on the correlation matrix). First we will consider analyses using the matrix of variances and covariances. A PCA generates m new variables, the principal components (PCs), by forming linear combinations of the original variables, $X = (X_1, X_2, \ldots, X_m)$, as follows:

$$PC_1 = b_{11}X_1 + b_{12}X_2 + \ldots + b_{1m}X_m = Xb_1$$

$$PC_2 = b_{21}X_1 + b_{22}X_2 + \ldots + b_{2m}X_m = Xb_2$$

$$\vdots$$

$$PC_m = b_{m1}X_1 + b_{m2}X_2 + \ldots + b_{mm}X_m = Xb_m \quad ,$$

where X_i have mean zero. In matrix notation,

$$P = (PC_1, PC_2, \ldots, PC_m) = X\,(b_1, b_2, \ldots, b_m) = XB,$$

and conversely $X = P\,B^{-1}$.

The rationale in the selection of the coefficients, b_{ij}, that define the linear combinations that are the PC_i is to try to capture as much of the variation in the original variables with as few PCs as possible. Since the variance of a linear combination of the Xs can be made arbitrarily large by selecting very large coefficients, the b_{ij} are constrained by convention so that the sum of squares of the coefficients for any PC is unity:

$$\Sigma_{j=1}^{m} b_{ij}^2 = 1 \qquad i = 1,2,\ldots,m \quad .$$

Under this constraint, the b_{1j} in PC_1 are chosen so that PC_1 has maximal variance.

If we denote the variance of X_i by s_i^2 and if we define the total variance as $T = \Sigma_{i=1}^{m} s_i^2$, then the proportion of the variance in the original variables that is captured in PC_1 can be quantified as $var(PC_1)/T$. In selecting the coefficients for PC_2, they are further constrained by the requirement that PC_2 be uncorrelated with PC_1. Subject to this constraint and the constraint that the squared coefficients sum to one, the coefficients b_{2j} are selected so as to maximize $var(PC_2)$. Further coefficients and PCs are selected in a similar manner, by requiring that a PC be uncorrelated with all PCs previously selected and then selecting the coefficients to maximize variance. In this manner, all the PCs are constructed so that they are uncorrelated and so that the first few PCs capture as much variance as possible. The coefficients also have the following interpretation which helps to relate the PCs back to the original variables. The correlation between the i^{th} PC and the j^{th} variable is

$$b_{ij}\sqrt{var(PC_i)}/s_j \quad .$$

After all m PCs have been constructed, the following identity holds:

$$var(PC_1) + var(PC_2) + \ldots + var(PC_m) = T = \Sigma_{i=1}^{m} s_i^2 \quad .$$

This equation has the interpretation that the PCs divide up the total variance of the Xs completely. It may happen that one or more of the last few PCs have variance zero. In such a case, all the variation in the data can be captured by fewer than m

variables. Actually, a much stronger result is also true; the PCs can also be used to reproduce the actual values of the Xs, not just their variance. We will demonstrate this more explicitly later.

The above properties of PCA are related to a matrix analysis of the variance-covariance matrix of the Xs, S_x. Let D be a diagonal matrix with entries being the eigenvalues, λ_i , of S_x arranged in order from largest to smallest. Then the following properties hold:

(i) $\lambda_i = \text{var}(PC_i)$

(ii) $\text{trace}(S_x) = \Sigma_{i=1}^{m} s_i^2 = T = \Sigma_{i=1}^{m} \lambda_i = \Sigma_{i=1}^{m} \text{var}(PC_i)$

(iii) $\text{corr}(PC_i, X_j) = \dfrac{b_{ij}\sqrt{\lambda}_i}{s_j}$

(iv) $S_x = B'DB$.

The statements made above are for the case when the analysis is performed on the variance-covariance matrix of the Xs. The correlation matrix could also be used, which is equivalent to performing a PCA on the variance-covariance matrix of the standardized variables,

$$Y_i = \frac{X_i - \bar{X}_i}{s_i}$$

PCA using the correlation martrix is different in these respects:

(i) The total "variance" is m, the number of variables.
(It is not truly variance anymore.)

(ii) The correlation between PC_i and X_j is given by $b_{ij}\sqrt{var(PC_i)} = b_{ij}\sqrt{\lambda_i} = \Lambda_i$ (called factor loading in SPSS-X). Thus PC_i is most highly correlated with the X_j having the largest coefficient in PC_i in absolute value.

The experimenter must choose whether to use standardized (PCA on a correlation matrix) or unstandardized coefficients (PCA on a variance-covariance matrix). The latter is used when the variables are measured on a comparable basis. This usually means that the variables must be in the same units and have roughly comparable variances. If the variables are measured in different units, then the analysis will usually be performed on the standardized scale, otherwise the analysis may only reflect the different scales of measurement. For example, if a number of fatty acid analyses are made, but the variances, s_i^2, and means, $\bar{X}_i$, are obtained on different bases and by different methods, then standardized variables could be used (PCA on the correlation matrix). To illustrate some of the above ideas, a number of examples have been constructed and these are described in Section 3. In each case, two variables, Z_1 and Z_2, which are uncorrelated, are used to construct X_i. Thus, all the variance can be captured with two variables and hence only two of the PCs will have nonzero variances. In matrix analysis terms, only two eigenvalues will be nonzero. An important thing to note is that in general, PCA will not recover the original variables Z_1 and

Z_2. Only standardized computations (PCA on correlation matrix) will be presented here because SPSS-X can only carry out a PCA based on a correlation matrix.

3. EXAMPLES

Throughout the examples we will use the variables Z_1 and Z_2 (with n = 11) from which we will construct $X_1, X_2, \ldots, X_m$. We will perform PCA on the Xs. Thus, in our constructed examples, there will only really be two underlying variables.

Values of Z_1 and Z_2

Z_1	-5	-4	-3	-2	-1	0	1	2	3	4	5
Z_2	15	6	-1	-6	-9	-10	-9	-6	-1	6	15

Notice that Z_1 exhibits a linear trend through the 11 samples and Z_2 exhibits a quadratic trend. They are also chosen to have mean zero and be uncorrelated. Z_1 and Z_2 have the following correlation matrix:

Correlation Matrix of Z_1 and Z_2

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

A correlation matrix always has unities along its diagonal and the correlation between the i^{th} variable and the j^{th} variable in the i^{th} row and j^{th} column.

Example 1: In this first example we analyze Z_1 and Z_2 as if they were the data. If PCA is performed on the correlation matrix,

then the SPSS-X output is as follows (SPSS-X control language for this example and all subsequent examples is in the appendix and the bold face print was typed on computer output to explain the calculation performed):

```
128 JAN 87    SPSS-X RELEASE 2.2  FOR IBM VM/CMS
 11:08:43     Cornell University, Ithaca, NY     IBM 3090      VM/SP CMS

0For  VM/SP CMS       Cornell University, Ithaca, NY   License Number 18599

 Use INFO OVERVIEW for more information on:
0* INCLUDE - To bring in command files              * Improvements in:
 * RENAME VARS - To rename variables                *   MANOVA
 * AUTORECODE - To recode strings as numbers        *   TABLES
 * Relinking Usercode                               *

    1  0  SET WIDTH=80
    2  DATA LIST LIST/Z1 Z2
    3  FACTOR VARIABLES=Z1 Z2/
    4   ANALYSIS=Z1 Z2/
    5   CRITERIA=FACTORS(2)/                                          } SPSS-X
    6   EXTRACTION=PC/                                                } Control
    7   PRINT=UNIVARIATE INITIAL CORRELATION EXTRACTION FSCORE/       Language
    8   ROTATION=NOROTATE/                                            (details
    9   SAVE REG(ALL PRIN)/                                         in appendix)
0THERE ARE   908928 BYTES OF MEMORY AVAILABLE.
 THE LARGEST CONTIGUOUS AREA HAS   908928 BYTES.
0THIS FACTOR ANALYSIS REQUIRES       1124 (       1.1K) BYTES OF MEMORY.
```

128 JAN 87 SPSS-X RELEASE 2.2 FOR IBM VM/CMS
11:08:44 Cornell University, Ithaca, NY IBM 3090 VM/SP CMS

0- - - - - - - - - - - F A C T O R A N A L Y S I S - - - - - - - - - - -

0
0ANALYSIS NUMBER 1 LISTWISE DELETION OF CASES WITH MISSING VALUES

	MEAN $\bar{Z}_i$	STD DEV S_i	LABEL
Z1	.00000	3.31662	
Z2	.00000	9.26283	

NUMBER OF CASES = 11

CORRELATION MATRIX: $= r_{ij}$

	Z1	Z2
Z1	$r_{11} = 1.00000$	
Z2	$r_{12} = r_{21} =$ .00000	1.00000 $= r_{22}$

EXTRACTION 1 FOR ANALYSIS 1, PRINCIPAL-COMPONENTS ANALYSIS (PC)

INITIAL STATISTICS:

VARIABLE	COMMUNALITY $= r^2_{Z_i;PC_1,PC_2}$	*	FACTOR $= PC_i$	EIGENVALUE $\lambda_i = S^2_{PC_i}$	PCT OF VAR	CUM PCT
Z1	1.00000	*	1	1.00000	50.0	50.0
Z2	1.00000	*	2	1.00000	50.0	100.0

0 PC EXTRACTED 2 FACTORS.

FACTOR MATRIX: = component loadings $= b_i\sqrt{\lambda_i} = \Lambda_i$

	FACTOR 1 $= \Lambda_1$	FACTOR 2 $= \Lambda_2$	$b_i = \Lambda_i/\sqrt{\lambda_i}$
Z1	.00000	1.00000	b_1 = [0 1]/1
Z2	1.00000	.00000	= [0 1]

- - - - - - - - - - - F A C T O R A N A L Y S I S - - - - - - - - - - -

FINAL STATISTICS: For PCA, "Final Statistics always is equivalent to "Initial Statistics" above.

VARIABLE	COMMUNALITY	*	FACTOR	EIGENVALUE	PCT OF VAR	CUM PCT
Z1	1.00000	*	1	1.00000	50.0	50.0
Z2	1.00000	*	2	1.00000	50.0	100.0

SKIPPING ROTATION 1 FOR EXTRACTION 1 IN ANALYSIS 1

FACTOR SCORE COEFFICIENT MATRIX: $= b_i/\sqrt{\lambda_i} = Y_i$

	FACTOR 1 $= Y_1$	FACTOR 2 $= Y_2$
Z1	.00000	1.00000
Z2	1.00000	.00000

These coefficients are used to compute the factor scores which appear on the next page

$$PC_i = Y_{i1}X_1/S_1 + Y_{i2}X_2/S_2$$

COVARIANCE MATRIX FOR ESTIMATED REGRESSION FACTOR SCORES:

$= S_{PC_iPC_j}$

	FACTOR 1	FACTOR 2
FACTOR 1	1.00000	
FACTOR 2	.00000	1.00000

```
128 JAN 87 SPSS-X RELEASE 2.2  FOR IBM VM/CMS
11:08:44  Cornell University, Ithaca, NY     IBM 3090      VM/SP CMS

0PRECEDING TASK REQUIRED   0.03 SECONDS CPU TIME;   0.64 SECONDS ELAPSED.

    11  PRINT /ALL
    12  EXECUTE
```

Z_1	Z_2	PC_1	PC_2
-5.00	15.00	1.61938	-1.50756
-4.00	6.00	.64775	-1.20605
-3.00	-1.00	-.10796	-.90453
-2.00	-6.00	-.64775	-.60302
-1.00	-9.00	-.97163	-.30151
.00	-10.00	-1.07958	.00000
1.00	-9.00	-.97163	.30151
2.00	-6.00	-.64775	.60302
3.00	-1.00	-.10796	.90453
4.00	6.00	.64775	1.20605
5.00	15.00	1.61938	1.50756

$PC_i = Y_{i1}X_1/S_1 + Y_{i2}X_2/S_2$

$PC_1 = 0X_1/3.32 + 1X_2/9.26$

for case 1,

$= 15/9.26$

$= 1.619$

The principal components are again the Xs (standardized Zs) themselves, but the eignvalues (var(PCs)) are unity since the variables have been standardized first.

Example 2: Let $X_1 = Z_1$, $X_2 = 2Z_1$ and $X_3 = Z_2$. SPSS-X gives the following results for a PCA on the correlation matrix:

```
128 JAN 87 SPSS-X RELEASE 2.2  FOR IBM VM/CMS
 11:09:55  Cornell University, Ithaca, NY     IBM 3090      VM/SP CMS

0- - - - - - - - - - -  F A C T O R  A N A L Y S I S  - - - - - - - - - - -

0
0ANALYSIS NUMBER  1  LISTWISE DELETION OF CASES WITH MISSING VALUES

                  MEAN      STD DEV   LABEL
                  X̄i          Si
 X1             .00000     3.31662
 X2             .00000     6.63325
 X3             .00000     9.26283

 NUMBER OF CASES  =       11


 CORRELATION MATRIX:  = rij

                 X1         X2         X3

 X1          1.00000
 X2          1.00000    1.00000
 X3           .00000     .00000    1.00000

0>WARNING 11302
 >The correlation matrix is ill-conditioned.


 EXTRACTION  1  FOR ANALYSIS  1, PRINCIPAL-COMPONENTS ANALYSIS (PC)


 INITIAL STATISTICS:

 VARIABLE      COMMUNALITY  *  FACTOR  EIGENVALUE    PCT OF VAR   CUM PCT
        = r²Xi;PC1,PC2,PC3  *    = PCi  = λi = s²PCi
 X1               1.00000  *       1      2.00000        66.7        66.7
 X2               1.00000  *       2      1.00000        33.3       100.0
 X3               1.00000  *       3       .00000          .0       100.0
0   PC EXTRACTED   3 FACTORS.
```

- - - - - - - - - - - F A C T O R A N A L Y S I S - - - - - - - - - - -

FACTOR MATRIX: = Component Loadings = $b_i\sqrt{\lambda_i} = \Lambda_i$

	FACTOR 1	FACTOR 2	FACTOR 3
X1	1.00000	.00000	.00000
X2	1.00000	.00000	.00000
X3	.00000	1.00000	.00000

FACTOR SCORE COEFFICIENT MATRIX: = $b_i/\sqrt{\lambda_i} = Y_i$

	FACTOR 1	FACTOR 2	FACTOR 3
X1	.50000	.00000	54794158.006
X2	.50000	.00000	-54794158.01
X3	.00000	1.00000	.00000

COVARIANCE MATRIX FOR ESTIMATED REGRESSION FACTOR SCORES: = $S_{PC_iPC_i}$

	FACTOR 1	FACTOR 2	FACTOR 3
FACTOR 1	1.00000		
FACTOR 2	.00000	1.00000	
FACTOR 3	.00000	.00000	1.00000

```
128 JAN 87 SPSS-X RELEASE 2.2  FOR IBM VM/CMS
 11:09:55  Cornell University, Ithaca, NY     IBM 3090      VM/SP CMS

 PRECEDING TASK REQUIRED   0.03 SECONDS CPU TIME;   0.54 SECONDS ELAPSED.

    12  PRINT /ALL
    13  EXECUTE
```

X_1	X_3	X_2	PC_1	PC_2	PC_3
-5.00	15.00	-10.00	-1.50756	1.61938	.00000
-4.00	6.00	-8.00	-1.20605	.64775	.00000
-3.00	-1.00	-6.00	-.90453	-.10796	.00000
-2.00	-6.00	-4.00	-.60302	-.64775	.00000
-1.00	-9.00	-2.00	-.30151	-.97163	.00000
.00	-10.00	.00	.00000	-1.07958	.00000
1.00	-9.00	2.00	.30151	-.97163	.00000
2.00	-6.00	4.00	.60302	-.64775	.00000
3.00	-1.00	6.00	.90453	-.10796	.00000
4.00	6.00	8.00	1.20605	.64775	.00000
5.00	15.00	10.00	1.50756	1.61938	.00000

$$PC_i = Y_{i1}X_1/S_1 + Y_{i2}X_2/S_2 + Y_{i3}X_3/S_3$$

$$PC_1 = .5X_1/3.317 + .5X_2/6.633 + 0X_3/9.26$$

for case 1,

$$= .5(-5)/3.317 + .5(-10)/6.633$$
$$= -1.50756$$

There are several items to note in these analyses:

i) There are only two nonzero eigenvalues since given X_1 and X_3, X_2 is computed from X_1.

ii) X_3 is its own principal component since it is uncorrelated with all the other variables.

iii) The sum of the eigenvalues is the number of variables, i.e.,

$$1 + 1 + 1 = 3 \ .$$

iv) Since there are only two nonzero eigenvalues, only two of the PCs have nonzero variances (are nonconstant).

v) The coefficients help to relate the variables and the PCs.

In the correlation analysis,

$$\text{Corr}(PC_1, X_1) = b_{11}\sqrt{\lambda_1} = \Lambda_{11} = \text{Component loading for } PC_1,\ X_1$$

$$= .707107\sqrt{2}$$

$$= 1 \ .$$

Thus, the variable is perfectly correlated with the PC.

vi) The Xs can be reconstructed exactly from the PCs with nonzero eigenvalues. For example, X_3 is clearly given by PC_2. X_1 and X_2 can be recovered via the formulas

$$X_1 = PC_1 \times S_1$$

$$X_2 = PC_1 \times S_2 \ .$$

As a numerical example,

$$\begin{aligned} -5 &= -1.508 \times 3.317 \\ -10 &= -1.508 \times 6.633 \end{aligned} \ .$$

Example 3: For Example 3 we use $X_1 = Z_1$, $X_2 = 2(Z_1+5)$, $X_3 = 3(Z_1+5)$ and $X_4 = Z_2$. Thus X_1, X_2 and X_3 are all created from Z_1. The data are:

OBS	X1	X2	X3	X4
1	-5	0	0	15
2	-4	2	3	6
3	-3	4	6	-1
4	-2	6	9	-6
5	-1	8	12	-9
6	0	10	15	-10
7	1	12	18	-9
8	2	14	21	-6
9	3	16	24	-1
10	4	18	27	6
11	5	20	30	15

The analysis for the correlation matrix (standardized analysis) is given below.

128 JAN 87 SPSS-X RELEASE 2.2 FOR IBM VM/CMS
11:10:27 Cornell University, Ithaca, NY IBM 3090 VM/SP CMS

0- - - - - - - - - - - F A C T O R A N A L Y S I S - - - - - - - - - - -

0
0ANALYSIS NUMBER 1 LISTWISE DELETION OF CASES WITH MISSING VALUES

	MEAN	STD DEV	LABEL
X1	.00000	3.31662	
X2	10.00000	6.63325	
X3	15.00000	9.94987	
X4	.00000	9.26283	

NUMBER OF CASES = 11

CORRELATION MATRIX: $= r_{ij}$

	X1	X2	X3	X4
X1	1.00000			
X2	1.00000	1.00000		
X3	1.00000	1.00000	1.00000	
X4	.00000	.00000	.00000	1.00000

INITIAL STATISTICS:

VARIABLE	COMMUNALITY $= r^2_{X_i;PC_1,PC_2,PC_3,PC_4}$	*	FACTOR $= PC_i$	EIGENVALUE $= \lambda_i = S_{PC_i}$	PCT OF VAR	CUM PCT
X1	1.00000	*	1	3.00000	75.0	75.0
X2	1.00000	*	2	1.00000	25.0	100.0
X3	1.00000	*	3	.00000	.0	100.0
X4	1.00000	*	4	.00000	.0	100.0

- - - - - - - - - - F A C T O R A N A L Y S I S - - - - - - - - - -

0 PC EXTRACTED 4 FACTORS.

FACTOR MATRIX: = Component Loadings = $b_i\sqrt{\lambda_i} = \Lambda_i$

	FACTOR 1	FACTOR 2	FACTOR 3	FACTOR 4
X1	1.00000	.00000	.00000	.00000
X2	1.00000	.00000	.00000	.00000
X3	1.00000	.00000	.00000	.00000
X4	.00000	1.00000	.00000	.00000

FACTOR SCORE COEFFICIENT MATRIX: = $b_i/\sqrt{\lambda_i} = Y_i$

	FACTOR 1	FACTOR 2	FACTOR 3	FACTOR 4
X1	.33333	.00000	46602552.201	.00000
X2	.33333	.00000	-42608047.73	.00000
X3	.33333	.00000	-3994504.474	.00000
X4	.00000	1.00000	.00000	.00000

COVARIANCE MATRIX FOR ESTIMATED REGRESSION FACTOR SCORES: = $S_{PC_iPC_j}$

	FACTOR 1	FACTOR 2	FACTOR 3	FACTOR 4
FACTOR 1	1.00000			
FACTOR 2	.00000	1.00000		
FACTOR 3	.00000	.00000	1.00000	
FACTOR 4	.00000	.00000	.00000	.00000

```
128 JAN 87 SPSS-X RELEASE 2.2  FOR IBM VM/CMS
 11:10:28  Cornell University, Ithaca, NY     IBM 3090      VM/SP CMS

0PRECEDING TASK REQUIRED   0.03 SECONDS CPU TIME;    0.58 SECONDS ELAPSED.

    13  PRINT /ALL
    14  EXECUTE
```

X_1	X_4	X_2	X_3	PC_1	PC_2	PC_3	PC_4
-5.00	15.00	.00	.00	-1.50756	1.61938	.00000	.00000
-4.00	6.00	2.00	3.00	-1.20605	.64775	.00000	.00000
-3.00	-1.00	4.00	6.00	-.90453	-.10796	.00000	.00000
-2.00	-6.00	6.00	9.00	-.60302	-.64775	.00000	.00000
-1.00	-9.00	8.00	12.00	-.30151	-.97163	.00000	.00000
.00	-10.00	10.00	15.00	.00000	-1.07958	.00000	.00000
1.00	-9.00	12.00	18.00	.30151	-.97163	.00000	.00000
2.00	-6.00	14.00	21.00	.60302	-.64775	.00000	.00000
3.00	-1.00	16.00	24.00	.90453	-.10796	.00000	.00000
4.00	6.00	18.00	27.00	1.20605	.64775	.00000	.00000
5.00	15.00	20.00	30.00	1.50756	1.61938	.00000	.00000

$$PC_i = Y_{i1}(X_1-\bar{X}_1)/S_1 + Y_{i2}(X_2-\bar{X}_2)S_2 + Y_{i3}(X_3-\bar{X}_3)/S_3 + Y_{i4}(X_4-\bar{X}_4)/S_4$$

$$PC_1 = .333(X_1-0)/3.317+.333(X_2-10)/6.633+.333(X_3-15)/9.950+0(X_4-0)/9.628$$

For case 1,

$$= .333(-5)/3.317 + .333(-10)/6.633 + .333(-15)/9.950$$
$$= -1.508$$

In the correlation analysis X_1, X_2 and X_3 have equal coefficients. As expected, the total variance is equal to the sum of the variances for the PCs and two PCs, PC_3 and PC_4, have zero variance and are identically zero.

Example 4. In this example we take more complicated combinations of Z_1 and Z_2.

$$X_1 = Z_1$$

$$X_2 = 2Z_1$$

$$X_3 = 3Z_1$$

$$X_4 = Z_1/2 + Z_2$$

$$X_5 = Z_1/4 + Z_2$$

$$X_6 = Z_1/8 + Z_2$$

$$X_7 = Z_2$$

Note that X_1, X_2 and X_3 are colinear (they all have correlation unity) and X_4, X_5, X_6 and X_7 have steadily decreasing correlations with X_1. The data are below.

OBS	X1	X2	X3	X4	X5	X6	X7
1	-5.000	-10.000	-15.000	12.500	13.750	14.375	15.000
2	-4.000	-8.000	-12.000	4.000	5.000	5.500	6.000
3	-3.000	-6.000	-9.000	-2.500	-1.750	-1.375	-1.000
4	-2.000	-4.000	-6.000	-7.000	-6.500	-6.250	-6.000
5	-1.000	-2.000	-3.000	-9.500	-9.250	-9.125	-9.000
6	0.000	0.000	0.000	-10.000	-10.000	-10.000	-10.000
7	1.000	2.000	3.000	-8.500	-8.755	-8.875	-9.000
8	2.000	4.000	6.000	-5.000	-5.500	-5.750	-6.000
9	3.000	6.000	9.000	0.500	-0.250	-0.625	-1.000
10	4.000	8.000	12.000	8.000	7.000	6.500	6.000
11	5.000	10.000	15.000	17.500	16.250	15.625	15.000

	X1	X2	X3	X4	X5	X6	X7

The PCAs for the and correlation matrix are given below.

128 JAN 87 SPSS-X RELEASE 2.2 FOR IBM VM/CMS
11:10:55 Cornell University, Ithaca, NY IBM 3090 VM/SP CMS

0- - - - - - - - - - - F A C T O R A N A L Y S I S - - - - - - - - - - -

ANALYSIS NUMBER 1 LISTWISE DELETION OF CASES WITH MISSING VALUES

	MEAN	STD DEV	LABEL
X1	.00000	3.31662	
X2	.00000	6.63325	
X3	.00000	9.94987	
X4	.00000	9.41010	
X5	.00000	9.29987	
X6	.00000	9.27210	
X7	.00000	9.26283	

NUMBER OF CASES = 11

CORRELATION MATRIX: $= r_{ij}$

	X1	X2	X3	X4	X5	X6	X7
X1	1.00000						
X2	1.00000	1.00000					
X3	1.00000	1.00000	1.00000				
X4	.17623	.17623	.17623	1.00000			
X5	.08916	.08916	.08916	.99614	1.00000		
X6	.04471	.04471	.04471	.99124	.99901	1.00000	
X7	.00000	.00000	.00000	.98435	.99602	.99900	1.00000

0- - - - - - - - - - - F A C T O R A N A L Y S I S - - - - - - - - - - -

INITIAL STATISTICS:

VARIABLE	COMMUNALITY	*	FACTOR	EIGENVALUE	PCT OF VAR	CUM PCT
		*				
X1	1.00000	*	1	4.05217	57.9	57.9
X2	1.00000	*	2	2.94783	42.1	100.0
X3	1.00000	*	3	.00000	.0	100.0
X4	1.00000	*	4	.00000	.0	100.0
X5	1.00000	*	5	.00000	.0	100.0
X6	1.00000	*	6	.00000	.0	100.0
X7	1.00000	*	7	.00000	.0	100.0

0 PC EXTRACTED 7 FACTORS.

FACTOR MATRIX: $= \Lambda_i$

	FACTOR 1	FACTOR 2	FACTOR 3	FACTOR 4	FACTOR 5
X1	.29046	.95689	.00000	.00000	.00000
X2	.29046	.95689	.00000	.00000	.00000
X3	.29046	.95689	.00000	.00000	.00000
X4	.99310	-.11729	.00000	.00000	.00000
X5	.97897	-.20399	.00000	.00000	.00000
X6	.96892	-.24739	.00000	.00000	.00000
X7	.95689	-.29046	.00000	.00000	.00000

	FACTOR 6	FACTOR 7
X1	.00000	.00000
X2	.00000	.00000
X3	.00000	.00000
X4	.00000	.00000
X5	.00000	.00000
X6	.00000	.00000
X7	.00000	.00000

- - - - - - - - - - - F A C T O R A N A L Y S I S - - - - - - - - - - -

FACTOR SCORE COEFFICIENT MATRIX: $= Y_i$

	FACTOR 1	FACTOR 2	FACTOR 3	FACTOR 4	FACTOR 5
X1	.07168	.32461	-32662298.18	76904163.301	1672138125.1
X2	.07168	.32461	10990731.797	-246283472.1	-583935891.6
X3	.07168	.32461	25793906.066	179875040.10	-1020020801
X4	.24508	-.03979	-16378276.30	-78251872.74	187767388.64
X5	.24159	-.06920	-5342035.794	6255117.6128	-417533700.2
X6	.23911	-.08392	-16992219.31	61205734.265	-1432368777
X7	.23614	-.09853	38417935.986	9652469.1703	1661978389.0

	FACTOR 6	FACTOR 7
X1	.00000	.00000
X2	.00000	.00000
X3	.00000	.00000
X4	.00000	.00000
X5	.00000	.00000
X6	.00000	.00000
X7	.00000	.00000

COVARIANCE MATRIX FOR ESTIMATED REGRESSION FACTOR SCORES: $= S_{PC_i PC_j}$

	FACTOR 1	FACTOR 2	FACTOR 3	FACTOR 4	FACTOR 5
FACTOR 1	1.00000				
FACTOR 2	.00000	1.00000			
FACTOR 3	.00000	.00000	1.00000		
FACTOR 4	.00000	.00000	.00000	1.00000	
FACTOR 5	.00000	.00000	.00000	.00000	1.00000

	FACTOR 1	FACTOR 2	FACTOR 3	FACTOR 4	FACTOR 5
FACTOR 6	.00000	.00000	.00000	.00000	.00000
FACTOR 7	.00000	.00000	.00000	.00000	.00000

	FACTOR 6	FACTOR 7
FACTOR 6	.00000	
FACTOR 7	.00000	.00000

```
16  PRINT /ALL
17  EXECUTE
```

X_1	X_7	X_2	X_3	X_4	X_5	X_6	PC_1
-5.00	15.00	-10.00	-15.00	12.50	13.75	14.38	1.11167
PC_2	PC_3	PC_4	PC_5	PC_6	PC_7		
-1.91293	.00000	.00000	.00000	.00000	.00000		
-4.00	6.00	-8.00	-12.00	4.00	5.00	5.50	.26951
-1.34220	.00000	.00000	.00000	.00000	.00000		
-3.00	-1.00	-6.00	-9.00	-2.50	-1.75	-1.38	-.36604
-.83418	.00000	.00000	.00000	.00000	.00000		
-2.00	-6.00	-4.00	-6.00	-7.00	-6.50	-6.25	-.79498
-.38888	.00000	.00000	.00000	.00000	.00000		
-1.00	-9.00	-2.00	-3.00	-9.50	-9.25	-9.13	-1.01731
-.00629	.00000	.00000	.00000	.00000	.00000		
.00	-10.00	.00	.00	-10.00	-10.00	-10.00	-1.03304
.31358	.00000	.00000	.00000	.00000	.00000		
1.00	-9.00	2.00	3.00	-8.50	-8.75	-8.88	-.84216
.57073	.00000	.00000	.00000	.00000	.00000		
2.00	-6.00	4.00	6.00	-5.00	-5.50	-5.75	-.44467
.76517	.00000	.00000	.00000	.00000	.00000		
3.00	-1.00	6.00	9.00	.50	-.25	-.63	.15943
.89689	.00000	.00000	.00000	.00000	.00000		
4.00	6.00	8.00	12.00	8.00	7.00	6.50	.97014
.96590	.00000	.00000	.00000	.00000	.00000		
5.00	15.00	10.00	15.00	17.50	16.25	15.63	1.98745
.97219	.00000	.00000	.00000	.00000	.00000		

Note: All variables could not fit on one line (width was set at 80 characters) so SPSS-X wraps around to next line to print remaining variables.

We note several things:

i) Only two eigenvalues are nonzero indicating that only two variables are needed. This is not readily apparent from the correlation or variance-covariance matrix.

ii) In PC_1, PC_2 and PC_3 where the standardizrd X_1, X_2 and X_3 are the same, they have the same coefficients.

iii) Neither PCA recovers Z_1 and Z_2. The PCs with nonzero variances have elements of both Z_1 and Z_2 in them, i.e., neither PC_1 or PC_2 is perfectly correlated with one of the Zs.

4. SUMMARY

PCA provides a method of extracting structure from the variance-covariance or correlation matrix. If a multivariate data set is actually constructed in a linear fashion from fewer variables, then PCA will discover that structure. PCA constructs linear combinations of the original data, X, with maximal variance:

$$P = XB \ .$$

In a correlation based PCA, as SPSS-X performs, linear combinations which maximize the variances of the standardized variables are constructed. This relationship can be inverted to recover the Xs from the PCs (actually only those PCs with nonzero eigenvalues are needed - see example 2). Though PCA will often help discover structure in a data set, it does have limitations. It will not necessarily recover the exact underlying variables, even if they were uncorrelated (Example 4). Also, by its construction, PCA is limited to searching for linear structures in the Xs.

APPENDIX

Control Language

Control language is typed in upper case and comments are in lower case. Refer to SPSS-X User's Guide, 1986, for program documentation.

Example 1

```
SET WIDTH=80
DATA LIST LIST/Z1 Z2  ⇒ Input data
FACTOR VARIABLES=Z1 Z2/
 ANALYSIS=Z1 Z2/       ⇒ Specifies which variables to use for PCA
 CRITERIA=FACTORS(2)/  ⇒ Specifies the number of factors or PCs to extract
 EXTRACTION=PC/        ⇒ Specifies a PCA
 PRINT=UNIVARIATE INITIAL CORRELATION EXTRACTION FSCORE/ ⇒ Specifies what
                                                          information is to
                                                          be printed
 ROTATION=NOROTATE/    ⇒ Instructs SPSS-X not to rotate the factors
 SAVE REG(ALL PRIN)/   ⇒ Instructs SPSS-X to compute factor scores for the
                         observations
BEGIN DATA
-5 15
-4 6
-3 -1
-2 -6
-1 -9
0 -10
1 -9
2 -6
3 -1
4 6
5 15
END DATA
PRINT /ALL            ⇒ causes the factor scores to be printed
EXECUTE
```

Example 2

```
SET WIDTH=80
DATA LIST LIST/X1 X3
COMPUTE X2=2*X1          ⇒ computes X2 from X1
FACTOR VARIABLES=X1 X2 X3/
 ANALYSIS=X1 X2 X3/
 CRITERIA=FACTORS(3)/
 EXTRACTION=PC/
 PRINT=UNIVARIATE INITIAL CORRELATION EXTRACTION FSCORE/
 ROTATION=NOROTATE/
 SAVE REG(ALL PRIN)/
BEGIN DATA
-5 15
-4 6
-3 -1
-2 -6
-1 -9
0 -10
1 -9
2 -6
3 -1
4 6
5 15
END DATA
PRINT /ALL
EXECUTE
```

Example 3

```
SET WIDTH=80
DATA LIST LIST/X1 X4
COMPUTE X2=2*(X1+5)
COMPUTE X3=3*(X1+5)
FACTOR VARIABLES=X1 X2 X3 X4/
 ANALYSIS=X1 X2 X3 X4/
 CRITERIA=FACTORS(4)/
 EXTRACTION=PC/
 PRINT=UNIVARIATE INITIAL CORRELATION EXTRACTION FSCORE/
 ROTATION=NOROTATE/
 SAVE REG(ALL PRIN)/
BEGIN DATA
-5 15
-4 6
-3 -1
-2 -6
-1 -9
0 -10
1 -9
2 -6
3 -1
4 6
5 15
END DATA
PRINT /ALL
EXECUTE
```

Example 4

```
SET WIDTH=80
DATA LIST LIST/X1 X7
COMPUTE X2=2*X1
COMPUTE X3=3*X1
COMPUTE X4=(X1/2)+X7
COMPUTE X5=(X1/4)+X7
COMPUTE X6=(X1/8)+X7
FACTOR VARIABLES=X1 X2 X3 X4 X5 X6 X7/
 ANALYSIS=X1 X2 X3 X4 X5 X6 X7/
 CRITERIA=FACTORS(7)/
 EXTRACTION=PC/
 PRINT=UNIVARIATE INITIAL CORRELATION EXTRACTION FSCORE/
 ROTATION=NOROTATE/
 SAVE REG(ALL PRIN)/
BEGIN DATA
-5 15
-4 6
-3 -1
-2 -6
-1 -9
0 -10
1 -9
2 -6
3 -1
4 6
5 15
END DATA
PRINT /ALL
EXECUTE
```